博弈中的反社会惩罚*

陈 璟[1,2] 张 融[#2] 袁佳琦[2] 余升翔[3]

（[1] 成都师范学院教育与心理学院，成都 611130）（[2] 四川师范大学心理学院，成都 610068）

（[3] 贵州财经大学工商管理学院，贵阳 550004）

摘 要 博弈中的反社会惩罚是指博弈者对表现出亲社会行为的高贡献或高合作性他人实施有代价经济惩罚、消极评价或排斥打压的现象。已有研究用带惩罚的经典博弈范式证明了反社会惩罚受到多种个体与环境因素的影响，并分别从侵犯、报复、社会比较、偏离群体规范、进化策略视角提出了解释其产生机制的五种假说。未来研究可在进一步厘清概念与测量指标、创新研究方法、拓展影响因素研究、明确产生机制、开展针对性干预研究方面做深入探索。

关键词 博弈，反社会惩罚，出色贬损，惩罚，合作

分类号

1 引言

古语有云：“人心齐，泰山移”。从古至今，合作一直是人类追求的终极目标之一，也是家庭、组织、社会乃至国家繁荣昌盛的基础。但现实生活中却存在诸多阻碍合作的因素，尤其是在利益相互制约的博弈情境中；例如，群体中的一些低贡献者通过孤立排斥、散播谣言等手段打压高贡献者、损毁其声誉。基于这类现象，研究者们开始了关于反社会惩罚的研究。早在 20世纪末，有研究者在社会困境经典研究范式公共物品博弈 (public goods game, PGG) 中引入惩罚机制，发现了社会困境中的私人惩罚现象 (也称非正式制裁 , informal sanctions , 即由非正式法律机构强加的、存在于私人之间的惩罚形式)(Falk et al., 2005; Fehr & Gächter, 2000) 。这种惩罚现象中既存在个体对破坏合作的搭便车者 (free-rider) 实施的利他惩罚 (altruistic punishment) (Fehr & Gächter, 2000; Gordon & Lea, 2016) ，又存在“有原则的搭便车者” (principled free rider) (Carpenter, 2007) 对高贡献合作者施加的恶意惩罚 (spiteful punishment) (Falk et al., 2005; Kirchkamp & Mill,

---

*收稿日期：2021-3-21
* 国家自然科学基金项目(71601136)、教育部人文社会科学研究项目(16YJC630005)、四川省科技计划项目软科学项目(2021JDR0349)和四川省省属高校科研创新团队建设计划资助。
共同第一作者
通信作者：余升翔，E-mail: shengxiangs@glut.edu.cn
陈璟，E-mail: cjbelinda@126.com

2020)。但多年以来，恶意惩罚因其难以被观测和量化、不具有类似利他惩罚的积极作用而被研究者所忽视（陈欣等, 2014; Herrmann et al., 2008）。直到2008年，Herrmann等人将恶意惩罚正式命名为反社会惩罚(antisocial punishment)，并通过包含16个国家样本的跨文化研究证明其广泛存在于不同文化背景的社会之中，其研究重要性才日益凸显。后续研究证实，反社会惩罚会严重削弱利他惩罚对合作的促进效果（陈思静，朱玥, 2020; 李晓博，马剑虹, 2017; Fatas et al., 2020)，危害团体的工作绩效，给强互惠理论带来巨大挑战（汪崇金，聂左玲, 2015; Kosfeld & Rustagi, 2015)，因此其研究价值不可小觑。

研究反社会惩罚中的人际互动模式，梳理与分析该领域的研究成果，有助于理解社会冲突与矛盾的个体心理根源并寻求干预途径，对于促进团体合作、强化组织治理、维护社会的和谐与稳定具有重要的理论意义和应用价值。然而，当前国外的该领域研究方兴未艾，我国则鲜见相关研究。基于此本文围绕反社会惩罚的概念与研究范式、影响因素和产生机制假说对该领域文献进行了系统梳理、分析与评价，提出了研究展望，以期抛砖引玉，激发国内学界对该主题的广泛关注。

2 反社会惩罚的概念与研究范式

2.1 反社会惩罚的概念及表现

惩罚如同双刃剑，与其“亲社会”一面（利他惩罚）相伴相生的是“反社会”的一面。命名者Herrmann等人(2008)将反社会惩罚定义为：博弈个体对表现出亲社会行为的其他博弈者的惩罚；并在PGG中以“惩罚者的贡献小于或等于被惩罚者”作为反社会惩罚的行为指标。上述概念和行为指标被后续的第二方惩罚研究（如：汪崇金等, 2018; Bryson et al., 2014; Sylwester et al., 2013) 广泛沿用。随着研究的发展，研究者发现反社会惩罚不只存在于第二方惩罚情境，不参与博弈的第三方也可能对做出公平分配的合作者进行惩罚(Morese et al., 2016)。在引入第三方之后，第二方惩罚研究对反社会惩罚的概念界定与行为指标均不再适用，于是有研究者以“被惩罚者在上轮博弈回合中的贡献额大于博弈参与者们的平均贡献额”为指标来界定反社会惩罚(Fehr & Williams, 2018)，并发现这一指标对合作的消极发展更具预测性(Fu & Putterman, 2018)；有研究者将以此为指标观察到的反社会惩罚称为反常惩罚(perverse punishment) (Ertan et al., 2009)。

反社会惩罚的表现形式具有多样性，除了典型的经济惩罚，还包括行为打压、诋毁排挤和消极评价等。例如：为诋毁高道德水平他人而实施出色贬损(do-gooder derogation) (Minson & Monin, 2012; Tasimi et al., 2015)；为阻碍团体变革、打压模范员工而对其实施同事消极约束（杜旌等, 2014)；将慷慨大方的高合作性成员踢出团体(Parks & Stone, 2010)；对表现越慷慨的人越不喜

欢 (Kawamura & Kusumi, 2020)，对其能力给予较低评价(Klein et al., 2015)。由这些多元化表现可知，前述经典概念已不能完全涵盖生活中的反社会惩罚现象。我们将博弈中的反社会惩罚定义为：博弈者对表现出亲社会行为的高贡献或高合作性他人实施有代价经济惩罚、消极评价或排斥打压的现象。

2.2 反社会惩罚的研究范式

反社会惩罚的研究范式包含带惩罚的公共物品博弈 (PGG)、囚徒困境博弈 (prisoner's dilemma, PD)、第三方惩罚博弈 (third party punishment game, TPPG)，以及联合独裁者博弈 (dictator game, DG) 进行评估的最后通牒博弈 (ultimatum game, UG)。

其中属 Herrmann 等人 (2008) 设计的 PGG 惩罚范式最具代表性，其程序如下：各持 20 代币初始资金的参与者 4 人一组进行 PGG；在捐赠阶段，无论自己是否捐赠，每人都能获得公共物品池中金额 0.4 倍的代币 (如各捐 20 代币，即各获 80×0.4=32 代币)，且主试会公布所有成员的捐赠情况；在惩罚阶段，每名成员均可使用一定数量 (0～10 个) 的代币以 1:3 的代价影响率 (cost-to-impact, C/I) (即支付 1 代币实施惩罚，被惩罚者损失 3 代币) 对其他成员进行惩罚。

PD 惩罚范式中的参与者在 PD 博弈中需选择合作、背叛或惩罚 (Wu et al., 2009)；参与者若选择合作，其收益 -1(对方 +2)；若选择背叛，其收益 +1(对方 -1)；若选择惩罚，其收益 -1(对方 -4)； PD 惩罚博弈的收益矩阵如表 1 所示 (若参与者 A 选合作， B 选背叛，那么 A 的总收益为 -2， B 为 3)。

表1 PD 惩罚博弈的收益矩阵

	合作	背叛	惩罚
合作	(1, 1)	(-2, 3)	(-5, 1)
背叛	(3, -2)	(0, 0)	(-3, -2)
惩罚	(1, -5)	(-2, -3)	(-5, -5)

TPPG 常与 DG 或 PD 结合使用，由不参与利益分配的第三方在观察他人博弈后选择是否对参与者实施有代价惩罚。研究者普遍认为引入第三方惩罚可有效促进合作 (Jordan et al., 2016)，但另有研究发现，在以 DG、PD 为基础的惩罚实验中，即使是利益不相干的第三方也可能实施反社会惩罚 (例如： Gerfo et al., 2019; Goette et al., 2012)。

联合 DG 进行评估的 UG 也被用于研究反社会惩罚 (二者的核心区别在于， DG 应答者没有拒绝权)。在 UG 中，共同分配一笔钱的博弈双方分别为提议者 (proposer) 和应答者 (responder)，先由前者提出金额分配方案，再由后者选择接受或拒绝该方案。若应答者选择接受，

则照此分配；若其选择拒绝，则两人都一无所得。研究者们普遍认为， UG 应答者对不公平要约的拒绝就是一种代价高昂的利他惩罚 (如： Henrich et al., 2006)。但也正是这些 UG 应答者在担任 DG 独裁者时做出了分给接受者 0 元的极端不公平分配；因此，研究者将同时具备上述两种行为的 UG 应答者的拒绝行为视为反社会惩罚 (Brañas-Garza et al., 2014) ，这种界定受到了广泛关注。

总的来说，已有实验范式各有利弊。PGG惩罚范式适合模拟多人参与的动态博弈过程，生态性较好，但人数多、过程复杂所滋生的无关变量（人际互动、群体规范等）使得实验过程较难控制。PD 惩罚范式仅涉及两方博弈，实验环境及收益计算相对简洁，但其中的反社会惩罚效应又仅能体现于动态博弈程序，较难观测。 TPPG 与其他博弈范式的结合使其更具灵活性，但目前仍主要用于利他惩罚研究，其对反社会惩罚研究的适用性和结果的可推广性有待进一步验证。 UG 的任务复杂性（涉及到双方的相对权利等）导致了无关变量的引入，使其中的反社会惩罚效应不够清晰，且联合 DG 进行评估的合理性与规范性有待进一步检验。

3 反社会惩罚的影响因素

作为一种复杂的社会行为，反社会惩罚受到多种个体与环境因素的影响。相关研究虽不甚丰富且深度有限，却颇具广度。

3.1 个体因素

3.1.1 生理因素

雄性激素睾酮是影响反社会惩罚的核心生理因素。它可能会通过增强杏仁核的反应性使个体表现出攻击性 (Carré et al., 2017) ，其作用也会被其他激素所调节。PGG研究发现，在低皮质醇水平下，高睾酮水平个体表现出更多的反社会惩罚，但这种效应并未在利他惩罚中出现 (Pfattheicher et al., 2014) 。这显示了反社会惩罚相较于利他惩罚对于睾酮影响的敏感性。

3.1.2 心理因素

认知是影响反社会惩罚的首要心理因素，虽然已有研究较为零散、亟待整合提炼，但从四个方面充分证明了认知因素的重要作用。首先，社会认知是影响反社会惩罚的心理基础；个体的法治意识公平意识、规范意识及信任感越强，则越少实施反社会惩罚 (汪崇金等 , 2018; Balliet & Van Lange, 2013; Hermann et al., 2008) 。其次，个体的主观意愿对反社会惩罚的影响主要表现在以下两方面：①个体会努力使自身决策与其主观意愿保持一致；当个体对自己与他人的未来关系表现出较强兴趣时，其反社会惩罚的实施力度会变小 (Horne & Irwin, 2016) 。例如，引入孤独者策略 (loner strategy; 即参与者可选择成为领取固定报酬而不参与博弈的“孤独者”) 的研究结果显

示，孤独者比非孤独者更少实施反社会惩罚 (Pleasant & Barclay, 2018)。②当个体的主观意愿受到违背时，所产生的认知不一致可能促使其做出更多的反社会惩罚。例如，有研究者在PGG中设置了自愿捐赠和强制捐赠两种情境并引入孤独者策略，结果显示，相较于自愿捐赠组的孤独者和两组非孤独者，强制捐赠组的孤独者实施了更多的反社会惩罚 (García & Traulsen, 2012; Hauser et al., 2014; Rand & Nowak, 2011)。再次，直觉系统的激活会使某些个体实施更多的反社会惩罚。研究发现，直觉系统的激活会使具有施虐倾向的个体实施更高频的反社会惩罚，而抑制直觉系统则使其反社会惩罚频率显著下降(Pfattheicher et al., 2017)。最后，工作记忆似乎也会影响反社会惩罚。有研究通过无关信息干扰个体的工作记忆，结果出现了反社会惩罚发生率上升(dos Santos et al., 2014)。

人格导致的个体差异也会影响反社会惩罚，尤其是一些与人格"阴暗面"有关的特质可能使个体更倾向于攻击和伤害他人。首先受到关注的是精神病态 (psychopathy)； Masui 等人 (2012) 发现高精神病态低家庭支持的个体会实施力度更大的反社会惩罚。于是，部分研究者开始关注包含马基雅维利主义 (Machiavellianism)、自恋 (narcissism)、精神病态三个维度的黑暗三人格 (dark triad personality) 对反社会惩罚的影响，但研究结果并不一致。例如，有研究发现上述三个维度得分标准化后的平均分能正向预测不同文化中的反社会惩罚发生率(Deutchman & Raihani, 2017)，但这种计分方式并不多见；另有研究则发现，这三个维度与反社会惩罚均无关联，不过认知启动能使高施虐倾向个体做出更多反社会惩罚，而对低施虐倾向个体无此影响 (Pfattheicher et al., 2017)。

3.2 环境因素

3.2.1 任务情境因素

任务情境是影响博弈决策的重要环境因素，其中颇受关注的因素首推代价影响率 (cost-to-impact ratio, C/I)。在经济领域的有代价惩罚中，研究者通常认为惩罚金额会随着 C/I 的提高而减少 (Anderson & Putterman, 2006)，当 C/I 高到一定程度时参与者就会因不划算而放弃惩罚。但相关研究结果却不太一致。部分研究发现 C/I 的增加的确抑制了反社会惩罚的发生。例如：相比 C/I≥1 时， C/I<1 时反社会惩罚频率更高(童婷, 2017)，而 C/I=1 时反社会惩罚则消失无踪 (Falk et al., 2005)；同样，计算机模拟的 PGG 惩罚研究发现，当公共物品池中捐赠金额的翻倍值 r(r>1) 较低时，在 C/I<1 的情况下，若 C/I 较低，反社会惩罚的发生率会大于利他惩罚，而一旦提高 C/I，利他惩罚又会占据主导，不实施利他惩罚的合作者会成为反社会惩罚的主要对象，从而使实施利他惩罚的合作者受到保护 (Szolnoki & Perc, 2017)。但也有研究发现反社会惩罚不受 C/I 的影响 (Carpenter, 2007; Egas & Riedl, 2008)。

信息公开性是另一重要的任务情境因素。在匿名情境中，不了解惩罚情况的参与者既不受他人前一轮决策的影响，也无需担心承担实施惩罚的后果；而公开情境会曝光惩罚情况，出于对结果的预期及对自身形象的维护，博弈者很可能会调整策略。研究证实，信息公开与否会影响博弈者的惩罚判断 (Denant-Boemont et al., 2007; Nikiforakis, 2008) ；信息透明的非匿名情境能有效抑制反社会惩罚的发生 (Hilbe & Traulsen, 2012)；增加信息公开的内容、程度及受众人数均能有效降低反社会惩罚的发生率 (Kamei & Putterman, 2015; 汪崇金等 , 2018)。还有研究发现，当信息公开且描述性规范 (即一个群体的典型行为) 强时，反社会惩罚力度最大；而当描述性规范弱时，信息公开与否对反社会惩罚无显著影响 (Horne & Irwin, 2016)。

此外，情境的竞争性是又一重要因素。竞争情境带来的资源紧张和高压会迫使个体更关注自身的相对收益；为获得更多的社会资源和更高的社会地位，贬损竞争对手成为一种常见的竞争策略，所以竞争情境中的反社会惩罚发生率远高于非竞争情境 (Pleasant & Barclay, 2018; Sylwester et al., 2013) 。

3.2.2 群体因素

群体因素对反社会惩罚的影响主要体现在两个方面：一是个体对内外群体的区别对待，二是团体决策与个人决策的差异。在个体对内外群体的区别对待方面，外群体的出现会对现有群体形成威胁导致二者间滋生敌意和侵略性，因此人们更倾向于保护内群体成员，而对外群体成员更苛刻，这就在增强群体内部合作的同时带来了群体间的反社会惩罚 (Bernhard et al., 2006; Bryson et al., 2014; Goette et al., 2012) 。例如，有研究要求意大利被试观看意大利人和中国人进行 DG 任务的视频并完成 TPPG ，结果显示：相较于视频中 DG 接受者为意大利人的内群体条件，在接受者为中国人的外群体条件下，被试在看到视频中的中国独裁者做出公平分配后，其会实施更多的反社会惩罚 (Morese et al., 2016) 。但也有研究发现意大利样本中的上述效应并不显著 (Gerfo et al., 2019) 。造成上述差异的原因可能在于：不同于 Morese 等人 (2016) 将惩罚者的初始金额设置为明显小于被惩罚者， Gerfo 等人 (2019) 将二者的初始金额设置为相等，后者的设计很可能缓和了群体间的竞争氛围。

在团体决策与个人决策的差异方面，因为团体决策的结果代表多人总的决策倾向，所以在进行团体决策时，个人决策因效能较弱而难以决定决策结果；若团体中实施反社会惩罚的个体较少，那么“进行反社会惩罚”的决策便极易被团体否决。因此，以团体为单位代替个人进行集中惩罚能够更好地维持合作 (Fehr & Williams, 2018; Gross et al., 2016)。例如，相较于以个人为单位参与PGG 惩罚任务，以 3 人一组的团体为单位会降低反社会惩罚的发生率 (Auerswald et al., 2018) ；且即

使以个人为单位参与惩罚决策，若将个人决策汇总后再由所有团体成员投票决定惩罚与否，也能降低反社会惩罚的发生率 (Pfattheicher et al., 2018)。

3.2.3 社会文化与发展因素

跨文化研究表明，反社会惩罚在不同国家普遍存在但表现出文化差异 (Bruhin et al., 2020; Herrmann et al., 2008; Klein et al., 2015; Lucas & Malki, 2018; Wu et al., 2009) 。其影响主要来自社会文化和社会发展水平两个方面。一方面，社会文化塑造了人们在思维方式、价值取向、群体规范等诸多方面的差异。不同文化背景下的价值取向和个体对规范的解释可能影响其反社会惩罚。例如，个体可能会将他人的高合作行为视为对群体规范的偏离，相较于具有个人主义文化背景的美国人具有集体主义文化背景的日本人对偏离群体规范行为的容忍度更低 (Gelfand et al., 2011) ，更易对偏离群体规范的行为实施反社会惩罚 (Kawamura & Kusumi, 2020) 。

另一方面，社会发展水平制约着国家的社会法治建设和民众受教育程度。社会发展水平高的国家往往具有更健全的社会法治、更高的民众受教育程度和主观幸福感，反社会惩罚的发生率也更低 (Herrmann et al., 2008; Stavrova et al., 2013) 。早期研究表明，反社会惩罚更常发生在社会不平等程度高 (高权力距离) 、个体与群体间联系强 (低个人主义) 、性别间差异淡化及不确定性规避很高的地区 (Hofstede, 2001) 。经典研究也发现，希腊、土耳其、前苏联和中东样本中的反社会惩罚发生率很高，而美国、澳大利亚、远东和欧洲西北部样本中的反社会惩罚发生率较低，这可能是由于民主程度和人均 GDP 与反社会惩罚发生率呈负相关 (Herrmann et al., 2008) 。几项后续研究结果验证了上述观点，例如：俄罗斯和罗马尼亚样本的反社会惩罚发生率分别高于瑞士和美国样本 (Ellingsen et al., 2012; Gächter & Herrmann, 2009)，研究者认为这可能源于俄罗斯和罗马尼亚具有相对较弱的法治规范和民主程度；意大利本地学生比中国留学生具有更高频的反社会惩罚研究者认为这可能源于意大利的法治规范更弱 (Rabellino et al., 2016)。

4 反社会惩罚的产生机制假说

反社会惩罚的产生机制迄今仍众说纷纭，研究者们从不同视角提出的五种假说各有优势、相互补充，丰富和加深了学界对反社会惩罚产生机制的认识。

4.1 侵犯假说

侵犯假说将反社会惩罚视为一种攻击行为 (Masui et al., 2012) ，认为其源于个体的恶意动机或内在负面特质。该假说获得了两方面研究证据的支持。一方面，行为研究证明了某些个体是基于自身恶意而非惩罚对象的表现与特点去实施反社会惩罚。例如，参与者即便对博弈对手的信息知之甚少也会做出惩罚决策 (Grechenig et al., 2010) ；有的参与者宁愿付出代价也要享受破坏他人财产的过

程 (Abbink & Herrmann, 2011) 。另一方面，生理与人格研究证明，个体的睾酮水平、黑暗三人格特质、施虐倾向与精神病态得分均与反社会惩罚行为密切相关 (Deutchman & Raihani, 2017; Masui et al., 2012; Pfattheicher et al., 2014, 2017) 。侵犯假说从惩罚者的个人特质出发，在一定程度上解释了个体反社会惩罚的内在根源，但因为仅关注内因而流于片面，忽视了个体社会行为形成的多端性。

4.2 报复假说

与内生的恶意不同，报复是个体基于“以牙还牙” (tit-for-tat, TFT) 策略对自身所受惩罚的回应。早期研究者提出，对其他博弈成员惩罚行为的报复可能是反社会惩罚的产生原因 (Fehr & Gächter, 2000; Thöni, 2014)。例如，在 PGG 中受到利他惩罚的搭便车者更可能在后续回合中报复性地对高贡献者实施反社会惩罚，且惩罚力度与其此前所受惩罚数量呈正相关 (Herrmann et al., 2008) 。一些上述观点的支持者甚至直接将反社会惩罚称为“反击惩罚” (counter-punishment) (如： Denant-Boemont et al., 2007; Kamei & Putterman, 2015) 。此外，惩罚的匿名性使搭便车者只能猜测所受惩罚来自高贡献者，所以这种报复也被称为“盲目报复” (blind revenge) (Fehr & Gächter, 2000) 。报复假说仅适用于解释个体在多轮动态博弈中的反社会惩罚，对于一回合的静态博弈则毫无解释力度；同时，缺乏整体性和多端性考量。

4.3 社会比较假说

社会比较理论指出，个体在上行道德比较时会将高道德水平他人视为自身威胁，从而产生道德自卑感 (moral inferiority) 、道德困惑 (moral confusion) 及可预期的道德谴责 (anticipated moral reproach) ，出现羡慕、轻蔑、嫉妒或敌意 (Monin, 2007) 。基于此，有研究者用社会比较解释个体对高声誉他人的出色贬损，指出反社会惩罚是个体为应对声誉和自我概念所面临的威胁而采取的防御手段 (Kuběna et al., 2014; Minson & Monin, 2012) 。支持该假说的行为研究证据包括为避免自己在群体内比较中显得很差，个体会降低对慷慨同伴的偏好 (Tasimi et al., 2015)，甚至将其踢出团体 (Parks & Stone, 2010) 。社会比较假说关注个体在社会关系中的感受和体验，以及由此产生的自我防御反应，从机能主义的角度看，它较好地解释了反社会惩罚的动力机制，是对前两项假说的有益补充。但该假说所强调的情绪因素的作用机制，迄今仍缺乏相关实验证据；此外，它也没有考虑到人际关系等外部因素的影响，仍显片面。

4.4 偏离群体规范假说

群体作用在多人博弈中十分重要，博弈群体构成的同时会形成相应的群体规范，个体可能会为了维持这种规范而实施反社会惩罚。根据规范从众理论 (theories of normative conformity) ，群体中的大多数人倾向做出的相同行为即是典型行为——描述性规范 (descriptive norm) ，而其他行

为则是不合群的非典型行为 (Abrams et al., 2000)；当群体中出现非典型行为时，群体成员会更多地惩罚那些具有非典型行为的个体 (Bellezza et al., 2014; Kawamura & Kusumi, 2020; Klein et al., 2015)。这意味着捐赠过多或过少的个体均会因偏离群体规范而遭受惩罚。Irwin 和 Horne(2013)据此提出：对描述性规范的偏离是反社会惩罚的产生根源，偏离群体规范的个体更易遭受反社会惩罚。他们通过控制 PGG 中其他成员的平均捐赠金额的离散程度来操纵描述性规范的强度，结果发现：描述性规范会影响个体对他人的惩罚决策，且描述性规范越强，反社会惩罚力度越大 (Horne & Irwin, 2016)。该假说聚焦外部因素，相较于前三种假说，为反社会惩罚的产生机制提供了新颖的解释视角；但它忽略了内部因素，且将反社会惩罚与利他惩罚的产生原因等同，从而使研究者无法据此区分这两种性质对立的社会行为的产生机制。

4.5 进化策略假说

Sylwester 等人 (2013) 提出反社会惩罚是一种进化策略，惩罚合作者可能只是一种获得优势的方式，尽管这是一种自私的行为，却因为有利于人类生存而在进化过程中被保留下来。

一方面，反社会惩罚是一种对个体有利的策略。生物市场理论 (biological-markets theory) 主张，人们在选择伙伴时，更乐于选择那些有高合作声誉、地位且有能力为自己带来利益的“最优伙伴” (Barclay, 2016; Gordon & Lea, 2016)。将合作伙伴关系作为竞争资源的研究进一步证明，在竞争条件下参与者会更多地使用反社会惩罚来减少他人的合作性，以凸显自己在合作市场中的相对市场价值，从而使其在与他人合作时占据更有利地位 (Pleasant & Barclay, 2018)；在这个意义上，反社会惩罚成为了具有适应意义的竞争策略。计算机仿真模拟的进化模型也表明，实施反社会惩罚有助于个体获得潜在利益，故该策略会因不断得到强化而逐渐成为主导策略 (Powers et al., 2012; Rand et al., 2010)。而且同一个体既可能实施利他惩罚也可能实施反社会惩罚 (Eriksson et al., 2014)，说明个体对策略的使用并不具有一贯性，个体会为了适应多变的环境而灵活地选择惩罚策略。

另一方面，反社会惩罚也是一种对群体有益的策略。当群体一致对外时，群体内相对适应的优势(relative fitness advantage) (如：团体凝聚力) 会被强化，伤害外群体的反社会惩罚由此在进化中得以保留 (Goette et al., 2012; Sylwester et al., 2013)。有研究通过创设额外奖金制造竞争情境，让被试与内外群体中的他人分别进行 TPPG-PD 任务，结果发现：群体间竞争的增加加剧了个体对外群体成员的反社会惩罚，同时增强了群体内部的合作 (Goette et al., 2012)。

进化策略假说立足于进化心理学的视角，强调反社会惩罚对于个体和群体的适应性价值，具有一定的创新意义。但目前尚无直接、充分的实证研究证据表明反社会惩罚的进化历程和它对适应性的影响；尽管部分研究者采用计算机仿真模拟去展示反社会惩罚的进化过程，但模型分析的结果缺乏相

应的行为数据支持，此类研究的生态效度仍有待进一步考量；可见，该理论尚有待后续研究做进一步检验。

总的来说，以上五种假说各有侧重，从个人与群体、社会和进化等视角分别对反社会惩罚的产生机制进行了解释。前两项假说从个体层面出发，对反社会惩罚产生机制的解释更强调人性的“阴暗面”；三、四项假说强调社会情境与社会关系对反社会惩罚的重要影响；第五项假说则另辟蹊径，强调反社会惩罚对个体和群体的有利作用。这些假说均局限于个体社会行为发端的某一个方面，割裂了人格、认知、情绪与行为之间的复杂联系。

5 总结与展望

自正式概念提出至今，反社会惩罚领域研究发展不过短短十余年；虽已涌现出一批有价值的研究成果，但在概念与测量指标、研究方法、影响因素、作用机制等方面仍存在一系列亟待解决的问题。

5.1 进一步厘清概念与测量指标

如前所述，已有研究存在聚焦经济博弈而忽视其他领域、着眼第二方惩罚而忽视第三方惩罚的不足，因此反社会惩罚的经典概念未能完全涵盖其多样化的现实表现。虽然我们基于文献分析给出了新的概念界定，但仍有必要通过后续研究进一步厘清和完善概念界定，明确其内涵与外延。

与此同时，反社会惩罚在经济领域内的测量指标也因不同研究者的衡量标准不同而存在差异。例如：在 PGG研究中，关注合作进化的研究者倾向于采用绝对标准，将反社会惩罚的测量指标定为贡献额低于均值的捐赠者对贡献额高于均值的捐赠者的惩罚（如： Pleasant & Barclay, 2018）。而关注惩罚动机等内在机制的研究者则倾向于采用相对标准，即无论博弈者的捐赠是否高于均值，只要被惩罚者比惩罚者捐赠得多，该惩罚就是反社会惩罚 （如： Herrmann et al., 2008）。有研究者曾采用这两种指标分别对反社会惩罚数据进行编码，结果发现：两种指标界定方式的重叠部分较多，但绝对指标比相对指标对合作的负向预测作用更大 (Fu & Putterman, 2018) 。可见，未来研究应着力于寻求统一的测量指标，力求提升该领域研究结论的可比性、可推广性和适用性。

5.2 进一步创新研究方法

已有反社会惩罚实验研究基本都在实验室条件下进行，缺乏能广泛模拟现实环境的多样化实验情境。有少数研究利用数学建模对惩罚过程进行计算机仿真模拟 （如： Han et al., 2019; Rand et al., 2010; Szolnoki & Perc, 2017)，但其研究成果在推广和应用方面颇受限制。而且外显的评价方法很易受到社会赞许效应的消极影响。未来可在研究方法层面做如下创新尝试：

第1 ，开发更具生态效度的生活决策情境模拟任务。反社会惩罚的研究范畴包括但不仅限于经济博弈领域，为在更广泛的真实生活情景中深入探索其内在机制与干预策略，未来研究应基于同事消

极约束、消极评价和诽谤等多样化表现开展反社会惩罚的非经济博弈类研究，开发涉及面广、生态效度高的反社会惩罚研究任务。例如：Gal 和 Rucker(2021) 在重要生活决策 (important life decision) 的研究中突破了传统的金钱博弈实验，让个体在与其息息相关的婚育、医疗、职业等领域进行模拟决策，提升了研究结果的可推广性。未来研究可借鉴其任务创新方式，开展采用生活决策情境模拟任务的反社会惩罚研究。

第二，开发能揭示参与者真实态度与动机的内隐联系测验 (implicit association test; IAT)，筛查具有反社会惩罚倾向的个体。有研究者运用 IAT 测量个体的内隐利他行为 (吴睿等, 2018)，证实其更能评估个体的真实态度。未来研究可将 IAT 引入反社会惩罚的测量中，从而弥补传统外显测量任务易受社会赞许效应影响的弊端，更好地甄别团体中的易实施反社会惩罚者。

5.3 进一步拓展影响因素研究

已有的影响因素研究虽成果颇丰，但仍存在部分研究结论不一致的问题。为了增强研究的应用价值、探索有效的干预策略，未来研究或可从以下几方面进行拓展研究。

第一，考察性别、年龄和样本类型等人口学变量的影响。在性别方面，一些已有研究仅选择男性样本 (如：Goette et al., 2012; Pfattheicher et al., 2014)，这就导致其研究结论的可靠性受到取样偏差的消极影响。如前所述，男性的较高睾酮水平意味着其通常比女性具有更高的攻击性，而这对于反社会惩罚很可能存在重要影响。可见，未来研究可以关注博弈个体的生理性别甚至性别角色类型对其反社会惩罚的影响。在年龄方面，个体的报复与攻击行为会随着年龄的增长而增加 (Zhang et al., 2017)，反社会惩罚很可能表现出年龄差异。但目前仅有 Tasimi 等人 (2015) 的研究揭示了儿童群体中的反社会惩罚现象，鲜有研究涉及不同年龄段群体的反社会惩罚状况与差异，后续或可深入开展反社会惩罚的跨年龄发展研究。在样本类型方面，现有研究多局限于大学生样本，但已有研究证明社会样本比大学生样本更易做出反社会惩罚 (Henrich et al., 2006)。可见，未来应对研究样本进行拓展，针对更广泛的社会样本 (如各类职业群体、特殊人格样本)、重要的组织群体 (如基层公职人员) 开展反社会惩罚研究，或许有利于提出更完善、更富针对性的干预策略。

第二，考察信任、社会支持等人际因素的重要影响。反社会惩罚植根于人际互动，由此足见人际因素对其影响的重要性，但关注此因素的直接相关研究非常少见。一方面，关于信任对反社会惩罚的影响，鲜有直接相关研究。一项采用问卷法与实验法共同考察人际信任与反社会惩罚关系的研究表明，个体的信任感越强，其反社会惩罚越少 (汪崇金等, 2018)。该研究虽然首次探讨了信任与反社会惩罚的关系，但其探索仅停留在相关层面；且该研究仅考察了惩罚者的初始信任，既未对信任水平进行操纵，也未测量即时决策情境中博弈者对其他参与者的信任度，因而该研究结果无法作为证明信任与反社会惩罚间存在因果关系的充分证据，也难以全面说明信任对反社会惩罚的影响。另有少数

决策相关研究发现，惩罚与合作的关系会受到社会信任水平的影响 (Balliet & Van Lange, 2013)；在那些社会信任水平低的国家，反社会惩罚出现频率较高 (Herrmann et al., 2008)；启动不信任 (distrust) 会让个体怀疑他人的行为意图，从而使其对自身的道德评估比对合作伙伴的道德评估更偏袒 (Weiss et al., 2018)；高人际信任能拉近社会距离、带来更多合作，从而增加利他惩罚 (Jordan et al., 2016; Weiss et al., 2021)。另一方面，社会支持也可能在反社会惩罚中起重要作用，因为有研究发现缺乏家庭支持可能助长个体的反社会惩罚 (Masui et al., 2012)，但社会支持的具体作用尚不明确。综上可知，未来研究可基于信任和社会支持等人际因素展开，深入探索它们对反社会惩罚的影响模式与机制。

第三，开展本土化研究以深入探索我国文化对反社会惩罚的塑造或抑制作用。经典研究早已揭示了反社会惩罚的跨文化一致性与量化差异（如：Herrmann et al., 2008)，即反社会惩罚在不同文化背景下受到社会规范、法治、经济等众多因素的影响。而我国虽具有集体主义文化背景，但近年来出现了个人主义价值逐渐盛行的现象（蔡华俭等, 2020)；且我国幅员辽阔、民族众多、不同区域之间也存在较大的文化与风俗差异。这些特色为反社会惩罚的本土化研究提供了广阔的发挥空间。例如，已有的本土化研究表明，南北方居住环境和饮食习惯均可能影响人们的思维方式 (Talhelm et al., 2014)，对辣味的喜爱往往伴随着更高的感觉寻求、冒险、攻击行为（傅于玲等, 2018)。那么，地域文化与食辣文化是否均会对反社会惩罚产生特殊影响，就是值得关注的有趣主题。可见，深入探究我国的多元文化对反社会惩罚的塑造或抑制作用，是在该领域开拓创新性研究的重要切入点。

5.4 进一步明确产生机制

如前所述，已有的反社会惩罚产生机制研究仍停留在理论思辨层面，且基本都属于单一视角的假说建构。未来研究或可在以下方面做进一步探索。

第一，从神经机制层面明确反社会惩罚的产生机制。已有研究绝大多数集中于行为层面，相关的认知神经机制研究屈指可数。例如，以内外群体为变量的 TPPG 功能性磁共振研究发现，个体在实施反社会惩罚时其奖赏网络中腹内侧前额叶 (ventromedial prefrontal cortex, VMPFC) 的活动显著增强 (Morese et al., 2016)。Gerfo 等人 (2019) 采用阳极经颅直流电刺激 (tDCS) 分别增强 VMPFC 与右侧颞顶联合区 (right temporoparietal junction, rTPJ) 的活性，发现激活心理化网络中的 rTPJ 会增加反社会惩罚的频率与力度；而激活 VMPFC 则不会。此外，社会比较假说指出，反社会惩罚是个体应对负性情绪体验的自我防御 (Minson & Monin, 2012)；但其所涉及的具体情绪成分依然不详，且鲜见考察情绪作用机制的相关研究，在未来深入开展该领域神经机制

研究的过程中，不但可以重点关注 VMPFC 和 rTPJ，还可结合经颅磁刺激 (TMS) 和事件相关电位 (event-related potential, ERP) 等技术、结合认知神经科学实验与行为实验做进一步探索。

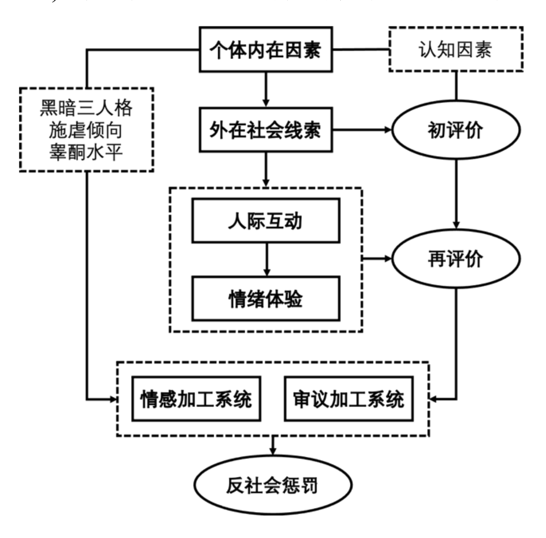


图1　反社会惩罚的双加工模型

第二，建构与检验解释反社会惩罚产生机制的心理模型。如前所述，已有的反社会惩罚产生机制理论模型均具片面性和缺乏整体性。因此，基于已有研究系统构建与检验解释反社会惩罚产生机制的心理模型十分必要。社会信息加工理论 (Crick & Dodge, 1994)主张，个体的行为是其对社会线索进行加工和解释的结果决策的双加工理· (dual-process theories) 主张，个份的决策是审议加工系统 (deliberative processes system) 与情感加工系统 (affective processes system) 相互作用的结果 (Evans, 2008) 。根据上述两个理论，综合考虑个体内在因素、外在社会线索和个体认知评价的重要作用，我们初步提出了反社会惩罚的双加工模型 (见图 1)，试图相对全面地解释反社会惩罚的产生机制；其主要观点是，反社会惩罚是个体基于自身内在因素，对外在社会线索、人际互动和情绪体验进行加工和解释的结果；针对不同的个体与决策情境，审议加工系统和情感加工系统分别被激活以主导决策行为。其具体过程可能如下：首先，博弈者的内在特质与相关激素水平能预测其实施反社会惩罚的可能。如博弈者在黑暗三人格、施虐倾向上有较高得分，或具有较高睾酮水平，则预示其实施反社会惩罚的可能性较大，此时其惩罚决策较少受外在社会线索的影响，直觉、情感等非理性因素的作用占主导，个体会更多基于情感加工系统的激活而做出反社会惩罚。若不具备上述前提，那么博弈者很可能会对其获取的社会规范、任务情境及其所受惩罚等外在情境中的社会线索进行初次评估；随着博弈的进行，个体会因与其他博弈参与者间的人际互动产生积极或消极的情绪体验。基于理性和情感的交互作用，个体将进行再次评估。若理性因素占优势，将激活审议加工系统，即个体基于理性权衡的结果做出决策；否则，将激活情感加工系统，即依据直觉或情感做出决策。

5.5 开展针对性干预研究

如前所述，反社会惩罚会给社会合作、团体绩效带来巨大危害，亟待深入开展具有针对性的干预研究。

第一，开展认知干预研究。已有研究从外在行为约束探讨了何种因素能够有效减少反社会惩罚的发生，如坚持信息公开(汪崇金等, 2018)、实施第三方惩罚(童婷, 2017; Gordon & Puurtinen, 2021; Zhou et al., 2017)、设立团体决策机制(Fehr & Williams, 2018)、提高惩罚代价(Falk et al., 2005)等，但并未涉及对个体内在因素的训练与提升研究。而已有研究发现，个体的规范感、信任感等社会认知水平越高，其反社会惩罚越少(汪崇金等, 2018)。可见，未来研究可以“开展认知干预以提升个体相关社会认知水平”为切入点，检验其对于反社会惩罚的干预效果。

第二，考察增添合作保障对于规避潜在反社会惩罚发生风险的作用。有研究在合作者策略选择中引入保险策略 (即合作者可选择预先付出一定代价购买“保险”以降低潜在风险，当遭受惩罚风险时便可获得保险补偿)，结果成功使合作重占优势，显著抑制了反社会惩罚 (张耀, 2016)。可见，为合作提供额外保障可能是有效的干预策略。例如，设置信用管理系统，当惩罚指向那些信用、声誉良好的个体时，便减轻惩罚的力度：或赋予这类个体购买“保险”降低受惩罚风险的特权，从而最大限度维护公平与合作。这一干预策略的效果有待未来研究的检验。

参考文献

蔡华俭, 黄梓航, 林莉, 张明杨, 王潇欧, 朱慧珺, 谢怡萍, 杨盈, 杨紫嫣, 敬一鸣. (2020). 半个多世纪来中国人的心理与行为变化——心理学视野下的研究. *心理科学进展*, *28*(10), 1599–1618.

陈思静, 朱玥. (2020). 惩罚的另一张面孔：惩罚的负面作用及破坏性惩罚. *心理科学*, *43*(4), 911–917.

陈欣, 赵国祥, 叶浩生. (2014). 公共物品困境中惩罚的形式与作用. *心理科学进展*, *22*(1), 160–170.

杜旌, 冉曼曼, 曹平. (2014). 中庸价值取向对员工变革行为的情景依存作用. *心理学报*, *46*(1), 113–124.

傅于玲, 邓富民, 杨帅, 徐玖平. (2018). 舌尖上的“自虐”——食辣中的心理学. *心理科学进*, *26*(9), 1651–1660.

童婷. (2017). *助人困境中反社会惩罚行为的影响因素及基于第三惩罚的干* (硕士学位论文). 华中师范大学, 武汉.

李晓博, 马剑虹. (2017). 公共物品两难中惩罚对合作的影响：偏好异质的视角. *社会科学家*, *6*, 67–71.

吴睿, 郭庆科, 李芳. (2018). 内隐和外显测量对利他行为的预测：来自 IAT 和 BIAT 的证据. *心理学探新*, *38*(4), 356–362.

汪崇金, 聂左玲. (2015). 破解社会合作难：强互惠真的够强?——基于公共品实验研. *外国经济与管理*, *37*(5), 53–65.

汪崇金, 史丹, 聂左玲, 崔凤. (2018). 打开天窗说亮话：社会合作何以可能. *中国工业经济*, *361*(04), 163–180.

张耀. (2016). *公共品博弈中合作演化及异质性研究* (硕士学位论文). 杭州电子科技大学.

Abbink, K., & Herrmann, B. (2011). The moral costs of nastiness. *Economic Inquiry, 49*(2), 631–633.

Abrams, D., Marques, J. M., Bown, N., & Henson, M. (2000). Pro-norm and anti-norm deviance within and between groups. *Journal of Personality and Social Psychology, 78*(5), 906–912.

Anderson, C. M., & Putterman, L. (2006). Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior, 54*(1), 1–24.

Auerswald, H., Schmidt, C., Thum, M., & Torsvik, G. (2018). Teams in a public goods experiment with punishment. *Journal of Behavioral and Experimental Economics, 72*, 28–39.

Barclay, P. (2016). Biological markets and the effects of partner choice on cooperation and friendship. *Current Opinion in Psychology, 7*, 33–38.

Balliet, D., & Van Lange, P. A. M. (2013). Trust, punishment, and cooperation across 18 societies: A meta-analysis. *Perspectives on Psychological Science, 8*(4), 363–379.

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature, 442*, 912–915.

Bellezza, S., Gino, F., & Keinan, A. (2014). The red sneakers effect: Inferring status and competence from signals of nonconformity. *Journal of Consumer Research, 41*(1), 35–54.

Brañas-Garza, P., Espín, A. M., Exadaktylos, F., & Herrmann, B. (2014). Fair and unfair punishers coexist in the ultimatum game. *Scientific Reports, 4*(1), Article 6025. http://dx.doi.org/10.1038/srep06025

Bruhin, A., Janizzi, K., & Thöni, C. (2020). Uncovering the heterogeneity behind cross-cultural variation in antisocial punishment. *Journal of Economic Behavior & Organization, 180*, 291–308.

Bryson, J. J., Mitchell, J., Powers, S. T., & Sylwester, K. (2014). Understanding and addressing cultural variation in costly antisocial punishment. In M. A. Gibson, & D. W. Lawson (Eds.), *Advances in the evolutionary analysis of human behaviour: Vol. 1: Applied evolutionary anthropology: Darwinian approaches to contemporary world issues* (pp. 201–222). New York, America: Springer.

Carpenter, J. P. (2007). The demand for punishment. *Journal of Economic Behavior & Organization, 62*, 522–542.

Carré, J. M., Geniole, S. N., Ortiz, T. L., Bird, B. M., Videto, A., & Bonin, P. L. (2017). Exogenous testosterone rapidly increases aggressive behavior in dominant and impulsive men. *Biological Psychiatry, 82*(4), 249–256.

Crick, N. R., & Dodge, K. A. (1994). A review and reformulation of social information-processing mechanisms in children's social adjustment. *Psychological Bulletin, 115*(1), 74–101.

Denant-Boemont, L., Masclet, D., & Noussair, C. N. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory, 33*, 145–167.

Deutchman, P., & Raihani, N. (2017). Dark Triad personality traits vary across countries and predict antisocial behavior.

dos Santos, M., Braithwaite, V. A., & Wedekind, C. (2014). Exposure to superfluous information reduces cooperation and increases antisocial punishment in reputation-based interactions. *Frontiers in Ecology and Evolution, 2*(36), 224–230.

Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society: Biological Sciences, 275*, 871–878.

Ellingsen, T., Herrmann, B., Nowak, M. N., Rand, D. G., & Tarnita, C. E. (2012). Civic capital in two cultures: The nature of cooperation in Romania and USA. *SSRN Electronic Journal, 89*(3), 575–581.

Eriksson, K., Cownden, D., Ehn, M., & Strimling, P. (2014). 'Altruistic' and 'antisocial' punishers are one and the same. *Social Science Electronic Publishing, 1*, 209–221.

Ertan, A., Page, T., & Putterman, L. (2009). Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *European Economic Review, 53*(5), 495–511.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*(1), 255–278.

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica, 73*(6), 2017–2030.

Fatas, E., Meléndez-Jiménez, M. A., & Solaz, H. (2020). Social hierarchies: A laboratory study on punishment patterns across networks. *Economic Inquiry, 58*(1), 104–119.

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review, 90*(4), 980–994.

Fehr, E., & Williams, T. (2018). Social norms, endogenous sorting and the culture of cooperation. *SSRN Electronic Journal*.

Fu, T., & Putterman, L. (2018). When is punishment harmful to cooperation? A note on antisocial and perverse punishment. *Journal of the Economic Science Association, 4*, 151–164.

Gächter, S., & Herrmann, B. (2009). Reciprocity, culture and human cooperation: Previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society of London, 364*(1518), 791-806.

Gal, D., & Rucker, D. D. (2021). Act boldly: Important life decisions, courage, and the motivated pursuit of risk. *Journal of Personality and Social Psychology, 120*(6), 1607–1620.

García, J., & Traulsen, A. (2012). Leaving the loners alone: evolution of cooperation in the presence of antisocial punishment. *Journal of Theoretical Biology, 307*(2), 168–173.

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., Duan, L., Almaliach, A., Ang, S., Arnadottir, J., Aycan, Z., Boehnke, K., Boski, P., Cabecinhas, R., Chan, D., Chhokar, J., D'Amato, A., Ferrer, M., Fischlmayr, I. C., ... Yamaguchi, S. (2011). Differences between tight and loose cultures: A 33-nation study. *Science, 332*(6033), 1100–1104.

Gerfo, E. L., Gallucci, A., Morese, R., Vergallito, A., & Lauro, L. J. R. (2019). The role of ventromedial prefrontal cortex and temporo-parietal junction in third-party punishment behavior. *NeuroImage, 200*, 501–510.

Goette, L., Huffman, D., Meier, S., & Sutter, M. (2012). Competition between organizational groups: Its impact on altruistic and antisocial motivations. *Management Science, 58*(5), 948–960.

Gordon, D., & Lea, S. E. G. (2016). Who punishes? The status of the punishers affects the perceived success of, and indirect benefits from, "moralistic" punishment. *Evolutionary Psychology, 14*(3), 1–14.

Gordon, D. S., & Puurtinen, M. (2021). High cooperation and welfare despite—and because of—the threat of antisocial punishments and feuds. *Journal of Experimental Psychology: General, 150*(7), 1373–1386.

Grechenig, K., Nicklisch, A., & Thöni, C. (2010). Punishment despite reasonable doubt—a public goods experiment with sanctions under uncertainty. *Journal of Empirical Legal Studies, 7*(4), 847–867.

Gross, J., Méder, Z. Z., Okamoto-Barth, S., & Riedl, A. (2016). Building the leviathan—voluntary centralisation of punishment power sustains cooperation in humans. *Scientific Reports, 6*(1), Article 20767. http://dx.doi.org/10.1038/srep20767

Han, D., Yan, S., & Li, D. (2019). The evolutionary public goods game model with punishment mechanism in an activity-driven network. *Chaos Solitons & Fractals, 123*, 254–259.

Hauser, O. P., Nowak, M. A., & Rand, D. G. (2014). Punishment does not promote cooperation under exploration dynamics when anti-social punishment is possible. *Journal of Theoretical Biology, 360*(25), 163–171.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2006). Costly punishment across human societies. *Science, 312*(5781), 1767–1770.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, 319*(5868), 1362–1367.

Hilbe, C., & Traulsen, A. (2012). Emergence of responsible sanctions without second order free riders, antisocial punishment or spite. *Scientific Reports, 2*(6), Article 458. http://dx.doi.org/10.1038/srep00458

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations*. Thousand Oaks, CA: Sage.

Horne, C., & Irwin, K. (2016). Metanorms and antisocial punishment. *Social Influence, 11*(1), 7–21.

Irwin, K., & Horne, C. (2013). A normative explanation of antisocial punishment. *Social Science Research, 42*(2), 562–570.

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature, 530*(7591), 473–476.

Kamei, K., & Putterman, L. (2015). In broad daylight: Fuller information and higher-order punishment opportunities can promote cooperation. *Journal of Economic Behavior & Organization, 120*, 145–159.

Kawamura, Y., & Kusumi, T. (2020). Altruism does not always lead to a good reputation: A normative explanation. *Journal of Experimental Social Psychology, 90*, Article 104021. https://doi.org/10.1016/j.jesp.2020.104021

Kirchkamp, O., & Mill, W. (2020). Conditional cooperation and the effect of punishment. *Journal of Economic Behavior and Organization, 174*, 150–172.

Klein, N., Grossmann, I., Uskul, A. K., Kraus, A. A., & Epley, N. (2015). It pays to be nice, but not really nice: Asymmetric reputations from prosociality across 7 countries. *Judgment & Decision Making, 10*(4), 355–364.

Kosfeld, M., & Rustagi, D. (2015). Leader punishment and cooperation in groups: Experimental field evidence from commons management in Ethiopia. *American Economic Review, 105*(2), 747–783.

Kuběna, A. A., Houdek, P., Lindová, J., Příplatová, L., & Flegr, J. (2014). Justine effect: Punishment of the unduly self-sacrificing cooperative individuals. *PLOS ONE, 9*(3), Article e92336. https://doi.org/10.1371/journal.pone.0092336

Lucas, P., & Malki, I. (2018). A note on the modelling and interpretation of a public goods game experiment. *Journal of Applied Statistics, 46*(4), 1–17.

Masui, K., Iriguchi, S., Terada, M., Nomura, M., & Ura, M. (2012). Lack of family support and psychopathy facilitates antisocial punishment behavior in college students. *Psychology, 3*(3), 284–288.

Minson, J. A., & Monin, B. (2012). Do-gooder derogation: Disparaging morally motivated minorities to defuse anticipated reproach. *Social Psychological & Personality Science, 3*(2), 200–207.

Monin, B. (2007). Holier than me? Threatening social comparison in the moral domain. *Revue Internationale De Psychologie Sociale, 2*0(1), 53–68.

Morese, R., Rabellino, D., Sambataro, F., Perussia, F., Valentini, M. C., Bara, B. G., & Bosco, F. M. (2016). Group membership modulates the neural circuitry underlying third party punishment. *PLOS ONE, 11*(11), Article e0166357. https://doi.org/10.1371/journal.pone.0166357

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of public economics, 92*(1-2), 91–112.

Parks, C. D., & Stone, A. B. (2010). The desire to expel unselfish members from the group. *Journal of Personality and Social Psychology, 99*(2), 303–310.

Pfattheicher, S., Böhm, R., & Kesberg, R. (2018). The advantage of democratic peer punishment in sustaining cooperation within groups. *Journal of Behavioral Decision Making, 31*(4), 562–571.

Pfattheicher, S., Keller, J., & Knezevic, G. (2017). Sadism, the intuitive system, and antisocial punishment in the public goods game. *Personality & Social Psychology Bulletin, 43*(3), 337–346.

Pfattheicher, S., Landhaeusser, A., & Keller, J. (2014). Individual differences in antisocial punishment in public goods situations: The interplay of cortisol with testosterone and dominance. *Journal of Behavioral Decision Making, 27*(4), 340–348.

Pleasant, A., & Barclay, P. (2018). Why hate the good guy? Antisocial punishment of high cooperators is greater when people compete to be chosen. *Psychological Science, 29*(6), 868–876.

Powers, S. T., Taylor, D. J., & Bryson, J. J. (2012). Punishment can promote defection in group-structured populations. *Journal of Theoretical Biology, 311*(4), 107–116.

Rabellino, D., Morese, R., Ciaramidaro, A., Bara, B. G., & Bosco, F. M. (2016). Third-party punishment: Altruistic and anti-social behaviours in in-group and out-group settings. *Journal of Cognitive Psychology, 28*(4), 486–495.

Rand, D. G., Armao, I. V., Joseph, J., Nakamaru, M., & Ohtsuki, H. (2010). Anti-social punishment can prevent the co-evolution of punishment and cooperation. *Journal of Theoretical Biology, 265*(4), 624–632.

Rand, D. G., & Nowak, M. A. (2011). The evolution of antisocial punishment in optional public goods games. *Nature Communication, 2*, Article 434. https://doi.org/10.1038/ncomms1442

Stavrova, O., Schlosser, T., & Fetchenhauer, D. (2013). Are virtuous people happy all around the world? Civic virtue, antisocial punishment, and subjective well-being across cultures. *Personality & Social Psychology Bulletin, 39*(7), 927–942.

Sylwester, H., Herrmann, B., & Bryson, J. J. (2013). Homo homini lupus? Explaining antisocial punishment. *Journal of Neuroscience, Psychology, and Economics, 6*(3), 167–188.

Szolnoki, A., & Perc, M. (2017). Second-order free-riding on antisocial punishment restores the effectiveness of prosocial punishment. *Physical Review X, 7*(4), Article 041027. https://doi.org/10.1103/PhysRevX.7.041027

Talhelm, T., Zhang, X., Oishi, S., Shimin, C., Duan, D., Lan, X., & Kitayama, S. (2014). Large-scale psychological differences within China explained by rice versus wheat agriculture. *Science, 344*(6184), 603–608.

Tasimi, A., Dominguez, A., & Wynn, K. (2015). Do-gooder derogation in children: The social costs of generosity. *Frontiers in Psychology, 6*, Article 1036. https://doi.org/10.3389/fpsyg.2015.01036

Thöni, C. (2014). Inequality aversion and antisocial punishment. *Theory & Decision, 76*(4), 529–545.

Weiss, A., Burgmer, P., & Mussweiler, T. (2018). Two-faced morality: Distrust promotes divergent moral standards for the self versus others. *Personality and Social Psychology Bulletin, 44*(12), 1712–1724.

Weiss, A., Michels, C., Burgmer, P., Mussweiler, T., Ockenfels, A., & Hofmann, W. (2021). Trust in everyday life. *Journal of Personality and Social Psychology, 121*(1), 95–114.

Wu, J. J., Zhang, B. Y., Zhou, Z. X., He, Q. Q., Zheng, X. D., Cressman, R., & Tao, Y. (2009). Costly punishment does not always increase cooperation. *Proceedings of the National Academy of Sciences,106*(41), 17448–17451.

Zhang, H., Qu, W., Ge, Y., Sun, X., & Zhang, K. (2017). Effect of personality traits, age and sex on aggressive driving: Psychometric adaptation of the Driver Aggression Indicators Scale in China. *Accident Analysis & Prevention, 103*, 29–36.

Zhou, Y., Jiao, P., & Zhang, Q. (2017). Second-party and third-party punishment in a public goods experiment. *Applied Economics Letters, 24*(1-3), 54–57.

Antisocial punishment in the game

CHEN Jing[1, 2], ZHANG Rong[2], YUAN Jiaqi[2], YU Shengxiang[3]

([1]*School of Education and Psychology, Chengdu Normal University, Chengdu 611130, China*)

([2]*School of Psychology, Sichuan Normal University, Chengdu 610068, China*)

([3]*School of Business Administration, Guizhou University of Finance and Economics, Guiyang 550004, China*)

Abstract: Antisocial punishment in the game refers to the phenomenon that the game participants implements economic punishment (with monetary cost), negative evaluation, or exclusion and suppression on others who exhibit high contributions or cooperation of prosocial behavior. Previous studies, using classic game paradigms with punishment, have proved that antisocial punishment is deeply influenced by a variety of individual and environmental factors, and put forward five hypotheses to explain its generation mechanism from the perspective of aggression, revenge, social comparison, deviation from group norms, and evolutionary strategies. Future researches can further clarify the concept and measurement indicators, innovate research methods, expand studies of influencing factors, explicate the generating mechanism, and conduct targeted intervention studies.

Key words: game, antisocial punishment, do-gooder derogation, punishment, cooperation